I Know That Person: Generative Full Body and Face De-Identification of People in Images

Karla Brkić[1] Ivan Sikirić[2] Tomislav Hrkać[1] Zoran Kalafatić[1]
[1]University of Zagreb, Faculty of Electrical Engineering and Computing, HR-10000 Zagreb, Croatia
[2]Mireo d.d., HR-10000 Zagreb, Croatia
karla.brkic@fer.hr ivan.sikiric@mireo.hr tomislav.hrkac@fer.hr zoran.kalafatic@fer.hr

Abstract

We propose a model for full body and face de-identification of humans in images. Given a segmentation of the human figure, our model generates a synthetic human image with an alternative appearance that looks natural and fits the segmentation outline. The model is usable with various levels of segmentation, from simple human figure blobs to complex garment-level segmentations. The level of detail in the de-identified output depends on the level of detail in the input segmentation. The model de-identifies not only primary biometric identifiers (e.g. the face), but also soft and non-biometric identifiers including clothing, hairstyle, etc. Quantitative and perceptual experiments indicate that our model produces de-identified outputs that thwart human and machine recognition, while preserving data utility and naturalness.

1. Introduction

Nowadays, the ever-increasing presence of cameras in daily life is taken for granted. From smartphones and CCTV to utility cameras in smart cars and smart home surveillance systems, we are used to being watched, photographed and have our data stored online. Photo and video sharing sites with dedicated smartphone apps make it as easy as one click for anyone to capture and upload potentially privacy-sensitive data to the cloud. Simultaneously, recent breakthroughs in deep learning have revealed a tremendous potential of computer vision in a wide array of real-world applications [17]. The community has been moving in leaps and bounds, making significant progress in solving very difficult problems including face recognition [25, 15], person re-identification [3, 34], image classification [11], etc. While these advances are immensely helpful in various applications that improve our daily lives, computer vision can very easily become an enabling technology for various attacks on privacy [28, 30, 23]. For example, computer vision can be used to connect real life identities to anonymous dating site profiles using Facebook profile photos [1]. Social media photos can be used to build sophisticated 3D models that can fool state of the art liveness and motion-enabled face recognition systems [35].

Figure 1. The effects of blurring for de-identification: while the faces are not recognizable, people can still be recognized from soft biometric and non-biometric identifiers such as hair color, clothing, hairstyle, personal items, etc. Images from the Clothing Co-Parsing (CCP) dataset [36].

Common attempts at thwarting unwanted identification of people in images and videos such as simple blurring, pixelization, etc., do very little to de-identify soft biometric and non-biometric identifiers [29] such as specifically colored and textured clothing, characteristic hairstyles and personal items, skin marks and tattoos, etc. (see Fig. 1). Despite a growing body of evidence suggesting that a wealth of information can be mined from publicly available photos even if the faces of humans are blurred, pixelated or even completely covered with a black box [21, 33, 23], commercial providers and general public still tend to view these naive transformations as an adequate form of privacy protection. Unfortunately, this late adoption of research results on de-identification seems to be a general trend [20]. Furthermore, visual quality and naturalness of the de-identified outputs are seldom taken into account.

In this work, our goal is to produce realistic de-identified

images of humans that thwart human and machine-based recognition. We introduce a model that de-identifies both primary biometric and often neglected soft biometric and non-biometric personal identifiers. Assuming that the silhouette of the person is known, we synthesize an alternative appearance that fits the silhouette and can therefore be seamlessly integrated into the original image. Our model is capable of producing results at varying level of detail, depending on the input resolution, output resolution requirements and the available segmentation of the human figure. If only a low resolution segmented human blob is available, e.g. obtained in a surveillance scenario using background subtraction, the model produces a low detail de-identified version of the blob. On the other hand, if a detailed segmentation at the garment level is available, the model produces a highly detailed de-identified version of the blob, considering each garment separately.

2. Related work

The majority of research on de-identification of humans in images still focuses on de-identifying primary biometric features, predominately the face. Earliest approaches to face de-identification involved applying naive transformations such as blurring or pixelization [30]. While naive transformations do thwart human recognition, it has been shown [8] that they can be effectively circumvented using computer vision, by employing a classifier trained on images on which the same transformation has been applied (so-called parrot recognition).

A higher level of privacy protection can be obtained through replacing the face with another, unrelated face, as e.g. in the work of Bitouk et al. [4]. They introduce a system that replaces faces by selecting a similar face from a database of real face images and adjusting the pose, lighting, and skin tone to match the original face. However, the construction of such a database for real applications is ethically and legally problematic, as rendering the faces of real people in de-identified sequences can potentially portray those people in a negative light. Although a number of large-scale face databases are available to the research community (e.g. MS-Celeb-1M [10], FaceScrub [22], CelebA [18] etc.), the data is licensed only for research purposes, with the disclaimers that the data providers do not own the data.

A viable alternative is to use synthetic face images for face replacement, as e.g. in the work of Newton et al. [21], who propose the k-same face de-identification algorithm. In the algorithm, each face image is replaced with an average of k most similar faces from a database where each person known to the system is represented with a single image. The algorithm is irreversible and offers a reasonable level of privacy protection, with the best possible success rate for re-identification being $1/k$. Several improvements to the k-same algorithm aim to improve data utility of the output, e.g. k-same select [7] and model-based k-same [9].

An important question to ask when considering identity protection through face de-identification is how much information can be mined on the individual even if the face is completely obfuscated. Oh et al. [23] propose a system for "faceless" person recognition and show that individuals in a social media setting can be identified even if their faces are blacked out and their clothing varies across images. Although a higher level of de-identification than face-only de-identification is needed, works on full-body de-identification are scarce. In an early work, Park and Trivedi [24] introduce a method for tracking and privacy protection in videos, with a de-identification scheme that covers individual human bounding boxes with colored rectangles. Agrawal and Narayanan [2] propose a method for tracking, segmenting and de-identifying individuals in videos. Two de-identification strategies are considered, one based on exponential pixel blurring and another based on line integral convolution. Although privacy-preserving, these schemes do not emphasize data utility and naturalness of images.

Our model for face and full body de-identification builds on recently introduced generative adversarial neural networks (GANs) [6]. GANs represent a promising new direction in artificial image synthesis, having enabled synthesizing images of various classes (e.g. faces, interiors, handwritten digits [6, 27]) that look highly realistic. In the GAN framework, two networks are trained, a generator and a discriminator. While the generator attempts to create an artificial image of the target class, the discriminator attempts to discern whether the image is artificial or real. Essentially, the two networks are engaged in a min-max game against each other. A number of extensions and improvements to GAN have been proposed, most notable the architecture known as deep convolutional generative adversarial network (DCGAN) [27] that introduces several architectural constraints and improves the stability of the training process. There have been works on applying GANs in a conditional setting, i.e. making the generated output conditioned over some input. For example, Zhu et al. [37] propose a method for user-guided visual manipulation of the generated images. Pathak et al. [26] introduce an inpainting technique that conditions the generated output that represents a missing image region to its surroundings. Isola et al. [14] propose a general image-to-image translation framework suitable for a variety of tasks including map-to-image translation, edge-to-image translation and image colorization.

Given the fact that deep networks have become a go-to solution for detection and recognition tasks in computer vision, the GAN framework is especially interesting in the context of de-identification, as it is tailored toward maximizing the potential of fooling a deep network (the discrim-

inator). By generating synthetic human images that fool the discriminator, we ensure not only that the generated humans look natural, but that a deep network cannot discern them from real images, making the re-identification harder.

3. Methodology

In this work, we assume that some kind of segmentation of the person to be de-identified is known (e.g. through using a person detector combined with a segmentation algorithm, or roughly estimated using background subtraction) and do not study how to obtain it. Rather, we propose a de-identification method that can work with segmentations of varying levels of detail: from simple foreground/background blob segmentations to precise garment-level annotations.

Our model builds on generative adversarial networks (GANs), attempting to generate synthetic samples from the distribution of all possible images that generated query segmentations. In general, the GAN framework consists of two parametric functions, a generator and a discriminator, modeled by deep neural networks. Given data $\mathbf{x}$, the generator maps its input variables $\mathbf{z}$ (noise with a prior $p_{\mathbf{z}}(\mathbf{z})$) to the data space using the mapping $G(\mathbf{z}, \theta_g)$, where θ_g are generator parameters. The discriminator $D(\mathbf{x}, \theta_d)$ with parameters θ_d takes data $\mathbf{x}$ as input and outputs a scalar that represents the probability that $\mathbf{x}$ was sampled from real data generating distribution p_{data}, and not from the generator distribution p_{g}. The two networks are trained simultaneously so that D maximizes the probability of correctly identifying generated samples, while G is attempting to minimize it. The discriminator and the generator are engaged in a min-max game with the objective function:

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}(\log D(\mathbf{x})) + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}(\log(1 - D(G(\mathbf{z})))). \quad (1)$$

The GAN objective is:

$$G^*_{\text{GAN}} = \arg\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D). \quad (2)$$

The GAN framework can be extended to model dependence between pairs of images, i.e. to make the generator output image conditioned on an input image, which is especially useful for various image to image translation tasks. Let $\mathbf{x}$ denote the input image, and let $\mathbf{y}$ be the image conditioned on $\mathbf{x}$ that we wish the generator to output. Both generator and discriminator take a pair of vectors as input: the discriminator takes the two mutually dependent images $\mathbf{x}$ and $\mathbf{y}$, while the generator takes the conditional image $\mathbf{x}$ and a noise vector $\mathbf{z}$. The objective function of a conditional GAN can be expressed as:

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{\text{data}}(\mathbf{x},\mathbf{y})}(\log D(\mathbf{x}, \mathbf{y})) + \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x}),\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}(\log(1 - D(\mathbf{x}, G(\mathbf{x}, \mathbf{z})))). \quad (3)$$

As shown in [14, 26], the quality of the images output from the generator can be improved by requiring them to be close to groundtruth images in accordance with some distance metric d:

$$\mathcal{L}_d(G) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{\text{data}}(\mathbf{x},\mathbf{y}),\mathbf{z}\sim p_{\mathbf{z}}}[d(G(\mathbf{x}, \mathbf{z}), \mathbf{y})]. \quad (4)$$

The conditional GAN objective can then be expressed as:

$$G^*_{\text{cGAN}} = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda\mathcal{L}_d(G). \quad (5)$$

3.1. Full body de-identification

In our model, we perform full body de-identification using a conditional GAN that generates synthetic human images. In order to ensure that the synthetic images look natural and fit in the original scene, the generative process is guided by the segmentations of the persons that are to be de-identified. The conditional GAN is trained on pairs of human segmentations and human images, with the goal of outputting realistically-looking synthetic human images conditioned on the extracted segmentation. Depending on the availability of data in particular applications, the network can be trained to operate on segmentations with varying levels of detail, from simple silhouette-defining blobs obtained e.g. using background subtraction to full body segmentations with detailed tags for individual garments.

3.1.1 Network architecture

The structure of our full body de-identification network is based on architectures described in [31, 27, 14].

Generator: The generator is an encoder-decoder network with skip connections as in the U-Net architecture [31]. There are eight encoder and eight decoder layers. The skip connections connect each encoder layer i with the decoder layer $n - i$, where n is the total number of layers (implemented as a concatenation of layer activations). Each layer consists of the following operations: convolution/deconvolution, batch normalization that normalizes the inputs to zero mean and unit variance [12], and leaky ReLU non-linearity [19]. We use filters of size 5×5 with a stride of 2 and zero padding of 2. In effect, each convolution downsamples the image by a factor of 2, while each deconvolution upsamples the image by a factor of 2.

Let `conv-n` denote a convolutional layer with n channels, batch normalization and leaky ReLU activations, `deconv-n` a deconvolution layer with n channels, batch normalization and leaky ReLU activations, and `deconv-dropout-n` a deconvolution layer with n channels, 50% dropout rate, batch normalization and leaky ReLU activations. The architectures of the encoder and the decoder are shown in Table 1.

Discriminator: The discriminator consists of four convolutional layers with batch normalization and leaky ReLU

Layer index	Encoder	Decoder
1	conv-64	deconv-dropout-512
2	conv-128	deconv-dropout-1024
3	conv-256	deconv-dropout-1024
4	conv-512	deconv-1024
5	conv-512	deconv-1024
6	conv-512	deconv-512
7	conv-512	deconv-256
8	conv-512	deconv-128

Table 1. The architecture of the conditional GAN generator (layer 8 of the encoder connects to layer 1 of the decoder).

activations organized as follows: conv64-conv128-conv256-conv512. The output of the last convolutional layer is transformed to a one-dimensional vector and a sigmoid function is applied to obtain the classification score. The filter size, stride and zero padding are the same as in generator layers, so in each convolutional layer the image is downsampled by a factor of 2.

Network training: To train the conditional GAN, we use the standard approach of Goodfellow et al. [6], i.e. mini-batch stochastic gradient descent training. In each iteration we perform one update of the discriminator, followed by one update to the generator. We use adaptive moment estimation (Adam) for gradient descent optimization [16].

3.2. Face de-identification

While in our model full body de-identification is performed using a conditional GAN, current state of the art conditional networks tend to produce outputs of a relatively low resolution and level of detail [37]. To improve the naturalness of the de-identified output, our model uses a dedicated GAN to synthesize artificial faces that are used as replacement faces on top of the de-identified full-body output. The face-generating GAN is not conditional, i.e. its output depends only on the random noise input $\mathbf{z}$ (see Eq. 1). The architecture of the face-generating GAN we use is the DCGAN of Radford et al. [27]. The architectural constraints are the same as for the conditional GAN we use for full body de-identification: only convolutional/deconvolution layers are used, along with batch normalization and ReLU activations; there are no pooling layers and no fully connected layers. Further details can be found in [27].

To determine the placement of the synthesized face on the full body de-identified image, we use a face detector on the original image. The synthesized face is then scaled to the size of the original face and blended with the full body de-identified image using an oval transparency mask.

4. Experimental evaluation

The experimental evaluation of our model is performed on two datasets, and it involves qualitative exploration, as well as perceptual and quantitative experiments.

Figure 2. Example images (top) and pixel-level segmentations (bottom) from the CCP dataset.

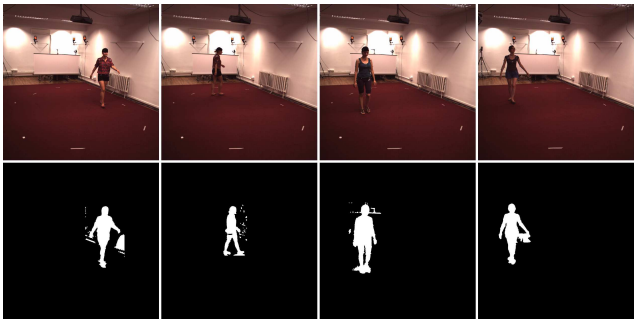

Figure 3. Example frames (top) and provided background subtraction segmentations (bottom) from the Human3.6M dataset.

4.1. Datasets

To study the effects of varying levels of the available input segmentation on the output de-identified images, we employ two datasets: the Clothing Co-Parsing (CCP) dataset [36] and the Human3.6M dataset [13]. The CCP dataset consists of 2098 street fashion images of a relatively high resolution (individual image size varies, the average being 828×550). Pixel-level segmentations of individual garments and skin and hair are available for 1004 images. There are a total of 59 segmentation tags defining various garment types, e.g. blazer, cardigan, sweatshirt, leggings, jeans, etc. A few example images from the CCP dataset and their corresponding segmentations are shown in Fig. 2

The Human3.6M dataset is a collection of 3.6 million video frames of professional actors performing various actions recorded in a controlled setting, with the corresponding 3D joint positions, laser scans of the actors and 3D human poses available. We use a subset of the dataset consisting of ten videos of different actors walking. The camera is static and the resolution of the videos is 1000×1002 pixels. Instead of pixel-level segmentations of actor parts, we use background subtraction masks available in the dataset as rough segmentation masks. A few example images and their corresponding background subtraction masks are shown in Fig. 3.

I_{orig}	I_{seg}	Network name
unchanged	garments	`clothing`
unchanged	blob only	`clothing-mono`
background masked	garments	`clothing-nobg`
background masked	blob only	`clothing-mono-nobg`

Table 2. Four training configurations of the conditional GAN for full-body de-identification.

4.2. Experimental setup

4.2.1 Full body de-identification

We train the conditional GAN for full body de-identification on the images from the CCP dataset for which pixel-level garment, skin and hair segmentations are available. There are a total of 1004 such images, and we use a train/test/validation split of 790/105/109 images. The groundtruth consists of pairs of images $(I_{\text{orig}}, I_{\text{seg}})$, where I_{orig} is the original image and I_{seg} is the corresponding segmentation on which the generator is to be conditioned. We consider two possibilities for supplying the original image: (i) supply the original image as is, and (ii) mask the background in the original image to prevent learning the background and increase network capacity for learning garment and body appearance. Additionally, we consider two possibilities for specifying the image I_{seg}: (i) supply the segmentation from the CCP dataset as is, with annotations for individual garments, and (ii) transform the CCP segmented image to a black and white blob that denotes only the silhouette of the person. This leads to a total of four network configurations summarized in Table 2.

Regarding the parameters of the training process, we follow the approach of [14] to introduce random jitter by resizing the training images to 286×286 and randomly cropping them to 256×256. Batch size is set to 1, leaky ReLU slope to 0.2, and we use L_1 as distance function (see Eq. 5). The parameter λ (Eq. 5) is set to 100.

4.2.2 Face synthesis

For face synthesis, we use the pretrained DCGAN model of [27], trained on their Faces dataset consisting of 10 million images of 10 thousand people. As the face detector needed for integrating the synthesized faces with the de-identified full body image we use the OpenCV implementation of the standard Viola-Jones detector [32, 5]. Given the original image and the full-body de-identified output, we detect the face in the original image, synthesize a random face using DCGAN and render the synthesized face on the location of the original face in the full-body de-identified output using an oval blending mask with a degree of transparency on the edges of the mask. Some examples of randomly synthesized faces are shown in Fig. 4.

Figure 4. Example faces synthesized by DCGAN.

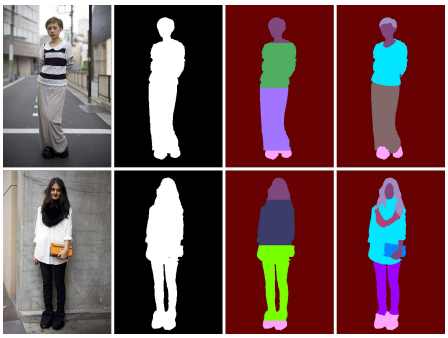

Figure 5. Two examples of naive segmentation. From left to right: original image, segmentation blob, naive segmentation algorithm output, groundtruth garment annotations.

4.3. Test sets and experiments

To test the performance of full body de-identification, we apply the four trained conditional GANs listed in Table 2 on their test splits. We term the test split of networks conditioned on full garment segmentations from CCP (`clothing`, `clothing-nobg`) as `clothing-garments`, and the test split of networks conditioned on person blobs from CCP (`clothing-mono`, `clothing-mono-nobg`) as `clothing-blobs`.

Additionally, we use a test set termed `bsblobs` consisting of background subtraction blobs from ten walking videos of the Human3.6M dataset. Ten background subtraction blobs of a person walking have been extracted per video, so the total size of the `bsblobs` set is 100 images. The `bsblobs` set is used to explore whether the trained networks generalize to other datasets, as well as to explore how the trained networks perform in the presence of noise in the segmentation input. Additionally, the setup of the Human3.6M dataset involving static cameras and background subtraction mimics a surveillance scenario, so it is especially interesting to investigate whether good de-identification can be achieved using simple background subtraction and applying our model.

4.3.1 Naive segmentation

In typical de-identification applications, garment-level segmentation of people is rarely available. Realistically, often the best we can hope for is that a blob outlining the

Experiment name	Network name	Test set
clothing	clothing	clothing-garments
clothing-nobg	clothing-nobg	clothing-garments
clothing-mono	clothing-mono	clothing-blobs
clothing-mono-nobg	clothing-mono-nobg	clothing-blobs
naiveseg	clothing-nobg	clothing-naiveseg
bsblobs	clothing-nobg	bsblobs
bsblobs-naiveseg	clothing-nobg	bsblobs-naiveseg

Table 3. Experimental setups for full body de-identification networks.

shape of the person's body is available, obtained e.g. using background subtraction. Aside from studying the performance of our conditional full body de-identification networks when initialized with such blobs only (the bsblobs set), we consider whether re-coloring such blobs using a naive segmentation strategy and applying networks trained on garment-level segmentations leads to better de-identified images.

Our naive segmentation algorithm is as follows: to separate the blob into four regions, depicting the head, top garment, bottom garment and shoes, the algorithm traverses the image pixel by pixel. The uppermost 10% of the blob area is denoted as the head region. The next 40% of the area is denoted as the top garment region, followed by the next 45% of the area denoting the bottom garment region and the lowermost 5% denoting the shoes/feet region. Once the regions have been assigned, region tags are randomly selected from the following pool of CCP tags: top = {*blazer, blouse, cardigan, hoodie, jacket, jumper, shirt, sweater, sweatshirt, t-shirt, top, vest*}, bottom = {*jeans, leggings, pants, shorts, tights*}. An example is shown in Fig. 5.

We apply our naive segmentation algorithm on the bsblobs set, and term the output set bsblobs-naiveseg. Additionally, we apply it on the clothing-blobs set, and term the output clothing-naiveseg.

4.4. Qualitative evaluation

To obtain a series of de-identified images and investigate the performance of individual defined full body de-identification networks (Table 2), the networks are applied on the described test sets in a total of seven experiments summarized in Table 3. Face synthesis is performed on all outputs to obtain final de-identified images.

Fig. 6 represents a qualitative exploration of different experiment outputs. In general, the level of detail of the input segmentation seems to determine the level of detail of the de-identified image. Note that the networks trained on datasets including backgrounds have also learned to produce synthetic backgrounds that look somewhat plausible, so the approach could have merit in location de-identification. Fig. 7 illustrates the performance on the bsblobs and bsblobs-naiveseg test sets. The outputs are of a notably lower quality than for segmentation inputs from the CCP dataset, showing that the networks ap-

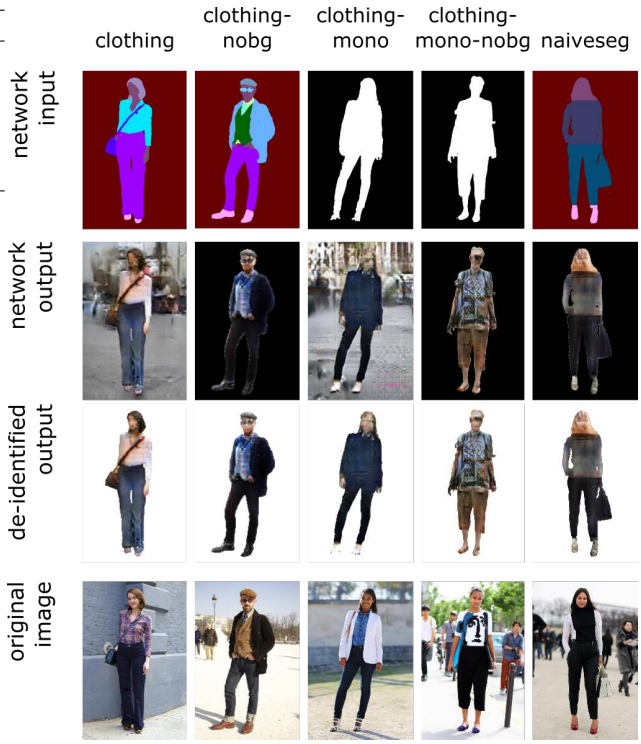

Figure 6. Visualization of the outputs of our de-identification model, depending on the used full body de-identification network and the level of detail in the input segmentation.

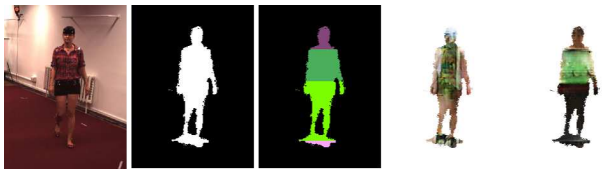

Figure 7. The outputs of our de-identification model on an example from the bsblobs set. From left to right: the original image, background subtraction blob, naive segmentation, output of the clothing-mono-nobg network using the background subtraction blob as input, and output of the clothing-nobg network using the naive segmentation as input.

Figure 8. The importance of independently synthesizing the face. From left to right: original image, de-identified full body image generated by the network clothing, a synthetic face added, output integrated into the original surroundings.

Experiment name	Garment colors and textures are similar [% yes]	Garment shapes are similar [% yes]	Mean naturalness score $\pm\sigma$	Mean recognizability score $\pm\sigma$
`clothing`	40.0	93.3	5.6 ± 2.6	3.9 ± 2.5
`clothing-nobg`	15.2	93.3	5.4 ± 2.6	4.2 ± 2.6
`clothing-mono`	13.3	32.4	3.6 ± 2.3	2.4 ± 1.7
`clothing-mono-nobg`	21.1	50.0	4.3 ± 2.1	2.8 ± 1.9
`naiveseg`	4.7	23.1	3.2 ± 2.1	2.0 ± 1.5
`bsblobs`	7.6	37.1	1.9 ± 1.3	1.5 ± 0.9
`bsblobs-naiveseg`	2.9	9.6	2.5 ± 2.1	1.5 ± 1.1

Table 4. Results of the perceptual study.

pear to be sensitive to noise, and that there is a cross-dataset loss in performance. Furthermore, the face detector failed on the majority of images in these datasets, contributing to the loss of naturalness.

A visualization of the merits of doing face synthesis as a separate step in our model is shown in Fig. 8. Note how the de-identified full body image lacks facial details, and how the naturalness is improved when a synthetic face is added.

As our training set includes images of real people, potentially generating images that closely resemble these people would present an ethical problem. However, we have found that the generated images do not resemble training set originals. Facial details in the generated images are lost (see Fig. 8), and garment information in the network is learned from many images, resulting in a diverse pool of potential renderings for each garment.

4.5. A perceptual study

Given a series of de-identified outputs obtained by experimental setups outlined in Table 3, we performed a perceptual study to quantitatively evaluate how humans perceive and respond to the de-identified images produced by our model. We randomly sampled five de-identified images from each of the seven experimental setups, resulting in a total of 35 images. The users were asked four questions about pairs of original and de-identified images: (i) whether the coloring and the textures of the garments on the two images are similar [yes/no], (ii) whether most garments are similarly shaped/tailored [yes/no], (iii) whether the de-identified image looks realistic, on a scale of 1-10 (1 - not realistic at all, 10 - completely realistic), (iv) to what degree the de-identified person is recognizable based on garments or face, on a scale of 1-10 (1 - not recognizable at all, 10 - completely recognizable). A total of 21 users participated in the study. Results are summarized in Table 4.

Ideally, our model should produce outputs that are perceptually distant from the originals (low garment color and texture similarity and low recognizability), while offering a high degree of naturalness. Although it is desirable for garment shape similarity to be low, we do not believe garment shape to be particularly identity-revealing, as a lot of people wear similarly tailored clothes. Overall, the highest mean naturalness score (5.6 on a scale of 1 to 10) is obtained for the `clothing` experimental setup (see Table 3) The second highest mean naturalness score (5.4) is obtained using a similar setup, `clothing-nobg`. Both `clothing` and `clothing-nobg` setups generate full body de-identifications with garment shapes similar as in the original image, which is to be expected given that the input segmentation exactly outlines individual garments. Interestingly, removing the background seems to have increased the capacity of the `clothing-nobg` network for varying garment colors and textures. Still, mean recognizability scores are approximately the same and do not reflect the perceived difference in garment colors and textures. We believe that these recognizability scores reflect the users' perceptual grouping according to the Gestalt principle of similarity: even though asked to score how well they could recognize the person based on the face and garment *colors and textures*, the users tended to consider the images similar, and therefore the person recognizable, if they viewed *garment shapes* as similar. This is supported by measuring the Pearson correlation coefficient between garment shape similarity (column 3 of Table 4) and mean recognizability score (column 5 of Table 4), which is a high 0.94.

Networks trained just on black and white blob segmentations (`clothing-mono` and `clothing-mono-nobg`) achieved lower mean naturalness score than their garment-trained counterparts. Mean recognizability scores were also lower. Removing the background of the original images yielded a higher naturalness score.

Applying the naive segmentation algorithm on the blobs from the CCP dataset (`naiveseg`) offered lower similarity of color, texture and shape and lower mean recognizability score compared to de-identification based on blobs only (`clothing-mono`), but at a cost of a slightly smaller mean naturalness score. Using naive segmentation on background subtraction blobs from the Human3.6M dataset (`bsblobs-naiveseg`) somewhat improved the naturalness of the de-identification compared to `bsblobs`. Overall, the naturalness score obtained on background subtraction blobs is quite low, which we attribute mainly to noisy

background subtraction input that generated noisy output (see Fig. 7) and to the fact that the face detector failed on over 90% of images from the Human3.6M dataset, so the de-identified outputs did not have faces rendered.

4.6. Re-identification performance

To quantitatively support our perceptual study findings, we perform a series of experiments aimed to mimic a simple re-identification attack scenario. We assume that the attacker has a gallery of images (in our case, we use all 2098 images from the CCP dataset), among which is the original image of the de-identified person, and study how likely it is that the person can be re-identified using the de-identified image as query (probe). We compare the performance for de-identified outputs of the first five of our seven experiments (i.e. we consider only the CCP dataset outputs, see Table 3) to performance for common naive de-identification techniques including (i) blurring the person ($\sigma = 10$), (ii) pixelization of the person, and (iii) covering the face with a black box and leaving the rest of the body as is. Each person is represented by a concatenation of a 3D histogram of RGB color components (20 bins per component) and a weighted gradient orientation histogram (20 bins). We measure k-nearest neighbor retrieval performance using the sum of histogram intersections for normalized color and gradient components as the distance measure.

Experimental results are summarized in Table 5. By applying our model, the likelihood of re-identification using color and gradient similarity is minimal even if the attacker has the exact same original image used to generate the de-identified image. In contrast, results for blurring, pixelization and black box-based re-identification support the intuition that individuals are still recognizable from cues other than face similarity.

Note that our representation does not consider silhouette shape, which could be used to trivially re-identify the de-identified images in this example, as our gallery images are the exact originals from which the de-identified images are generated. However, we assume that in real applications it is unlikely that the attacker has the original images.

4.7. Discussion

Applying our model on full body segmentations of any level of detail, from simple blobs and naively segmented blobs to full garment-based segmentations, produces de-identified outputs that are distant from the original images in terms of color and gradient similarity, as illustrated by our re-identification experiments. Simultaneously, our perceptual study shows that the the more detailed the segmentation input, the higher the naturalness of the de-identified output. However, naturalness and recognizability are also highly correlated (Pearson correlation coefficient of 0.97). As noted earlier, we believe that recognizability scores were influenced by garment shape similarity between pairs of original and de-identified images (correlation coefficient of 0.94). In real scenarios, it is our intuition that garment shape is not in itself as identifying as garment colors or textures, but this intuition remains to be corroborated in further experiments. As the users have recognized the de-identified garment shapes as similar to the real-world shapes in the original images, we believe that the lack of complete naturalness perceived by users is due to the de-identified colors and textures. As noted by Zhu et al. [37], GANs still have limitations (low resolution of generated results, blur and lack of texture), and they perform better on structured datasets than on more general imagery. We expect further improvements to the GAN state of the art will drive up the naturalness of the images generated by our model.

Experiment name	Accuracy [%]			
	$k=1$	$k=3$	$k=5$	$k=10$
`clothing`	0.0	1.1	3.2	3.2
`clothing-nobg`	1.1	1.1	1.1	2.1
`clothing-mono`	0.0	0.0	2.1	2.1
`clothing-mono-nobg`	0.0	0.0	0.0	1.1
`naiveseg`	0.0	0.0	0.0	0.0
`blur`	37.9	48.4	56.8	62.1
`pixelized`	87.4	97.9	97.9	100.0
`black-rect`	100.0	100.0	100.0	100.0

Table 5. Re-identification k-nn retrieval performance.

As several experimental setups and network training strategies were tried out (see Tables 2 and 3), our summary recommendation is to use as detailed input segmentation as possible when generating the de-identified images. When only noisy blobs are available, it might be beneficial in terms of output naturalness to smooth them as much as possible and to employ a naive segmentation strategy.

5. Conclusion

We have introduced a model for full body and face de-identification of humans in images that enables synthesizing artificial human images that fit an input segmentation. We have shown that the model is useful for various kinds of segmentations, from simple noisy bounding subtraction blobs to highly detailed garment-level segmentations. In a perceptual study, we have found that our model generates images that look reasonably natural (best-performing setup scored a 5.6 naturalness score on a scale of 1 to 10), while offering a solid level of identity protection. Quantitative exploration of de-identification performance found that the de-identified images produced by our model share very little color and gradient similarity with the original images, in contrast to naive methods including blurring, pixelization and covering the face with a black rectangle.

References

[1] A. Acquisti, R. Gross, and F. Stutzman. Faces of Facebook: Privacy in the age of augmented reality, August 2011. Black-hat USA Technical Security Conference . 1

[2] P. Agrawal and P. J. Narayanan. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(3):299–310, March 2011. 2

[3] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[4] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph.*, 27(3):39:1–39:8, 2008. 2

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 5

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. Conf. on Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014. 2, 4

[7] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. *Privacy Enhancing Technologies: 5th International Workshop*, chapter Integrating Utility into Face De-identification, pages 227–242. Springer Berlin Heidelberg, 2006. 2

[8] R. Gross, L. Sweeney, J. F. Cohn, F. De la Torre, and S. Baker. *Protecting Privacy in Video Surveillance*, chapter Face De-identification, pages 129–146. Springer Publishing Company Incorporated, 2009. 2

[9] R. Gross, L. Sweeney, F. de la Torre, and S. Baker. Model-based face de-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 161–161, 2006. 2

[10] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. *MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition*, pages 87–102. Springer International Publishing, Cham, 2016. 2

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. R. Bach and D. M. Blei, editors, *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015. 3

[13] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 4

[14] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. 2, 3, 5

[15] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[17] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 5 2015. 1

[18] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. Int. Conf. on Computer Vision (ICCV)*, 2015. 2

[19] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Frnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814. Omnipress, 2010. 3

[20] A. Narayanan, J. Huey, and E. W. Felten. *A Precautionary Approach to Big Data Privacy*, pages 357–385. Springer Netherlands, Dordrecht, 2016. 1

[21] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, Feb 2005. 1, 2

[22] H. W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *Proc. Int. Conf. on Image Processing (ICIP)*, pages 343–347, 2014. 2

[23] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Faceless person recognition; privacy implications in social media. *CoRR*, abs/1607.08438, 2016. 1, 2

[24] S. Park and M. M. Trivedi. A track-based human movement analysis and privacy protection system adaptive to environmental contexts. In *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*, pages 171–176, Sept 2005. 2

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015. 1

[26] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. 2016. 2, 3

[27] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 2, 3, 4, 5

[28] H. A. Rashwan, A. Solanas, D. Puig, and A. Martínez-Ballesté. Understanding trust in privacy-aware video surveillance systems. *International Journal of Information Security*, 15(3):225–234, 2016. 1

[29] D. Reid, S. Samangooei, C. Chen, M. Nixon, and A. Ross. Soft biometrics for surveillance: an overview. In *Machine Learning: Theory and Applications*, 31, pages 327–352. Elsevier, 2013. 1

[30] S. Ribarić, A. Ariyaeeinia, and N. Pavešić. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47:131 – 151, 2016. 1, 2

[31] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pages 234–241. Springer International Publishing, Cham, 2015. 3

[32] P. Viola and M. Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001. 5

[33] M. J. Wilber, V. Shmatikov, and S. J. Belongie. Can we still avoid automatic face detection? *CoRR*, abs/1602.04504, 2016. 1

[34] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[35] Y. Xu, T. Price, J.-M. Frahm, and F. Monrose. Virtual u: Defeating face liveness detection by building virtual models from your public photos. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 497–512, Austin, TX, 2016. USENIX Association. 1

[36] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 3182–3189, Washington, DC, USA, 2014. IEEE Computer Society. 1, 4

[37] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 2, 4, 8